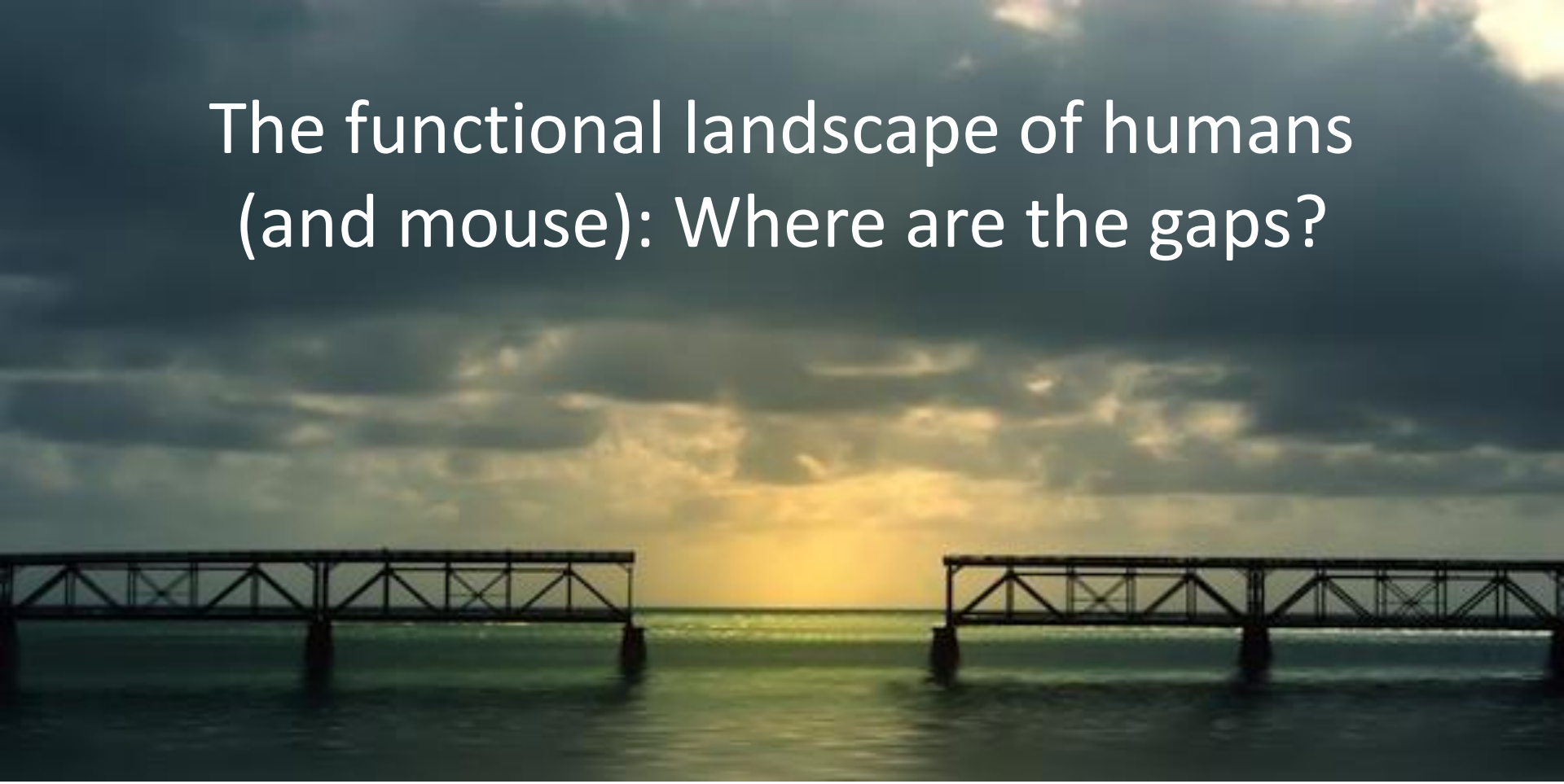


The functional landscape of humans (and mouse): Where are the gaps?

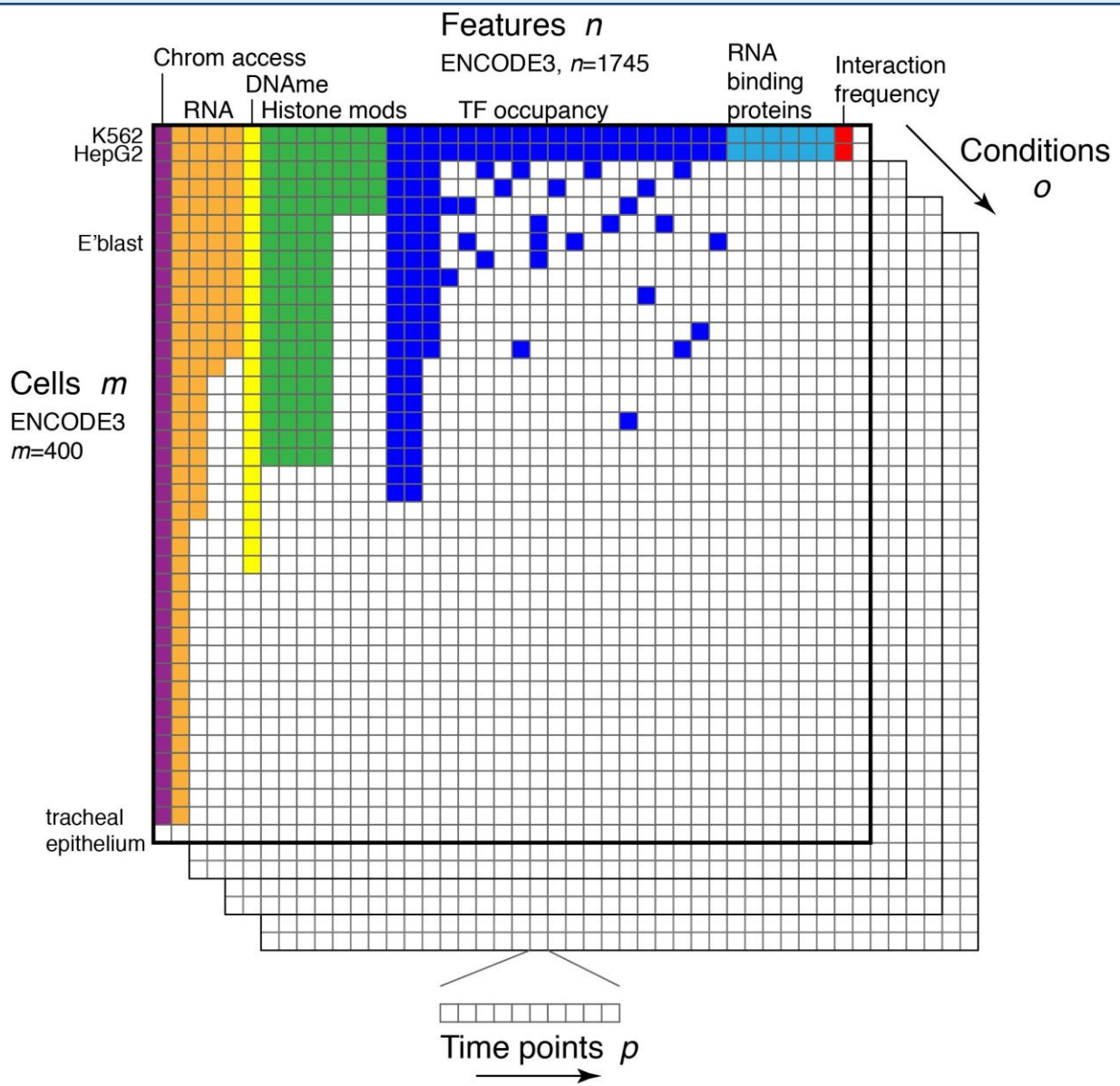
Ross Hardison

NHGRI Workshop: From functional genomics to biomedical insights, Mar 10-11, 2015

Questions from NHGRI and planners

- What is the current status of mapping functional elements in human and mouse?
- What high throughput, genome-wide, unbiased data production efforts are of highest priority?
- What data validation and characterization efforts should be undertaken?
- What future studies should be envisaged if not limited by technology?
- What technological breakthroughs would be transformative?
- How would you prioritize needs?
- What is needed for making new data interoperable with previous ENCODE findings?

The current state of mapping function-associated features

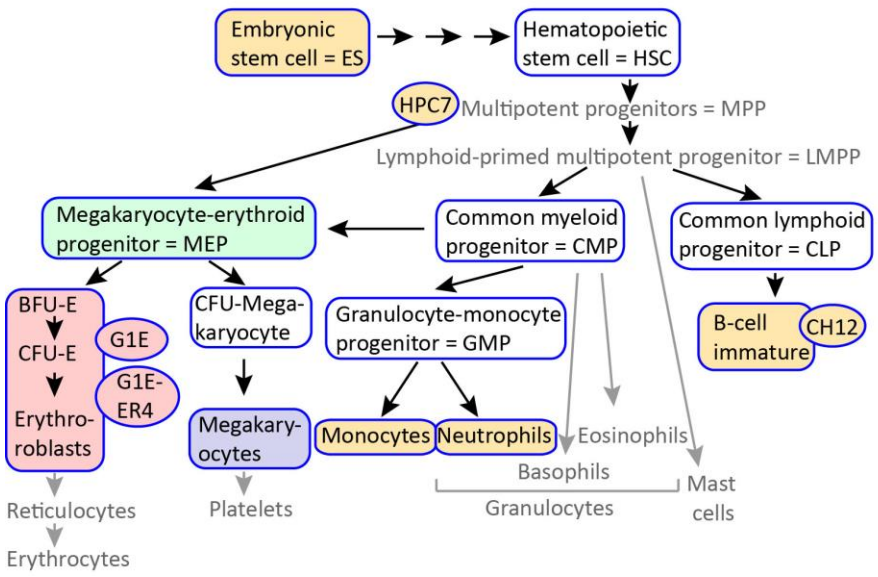


Brute force approach to completion

- Number of cell types m = 2000 (COPE database)
- Number of features n = 2000 (ca 1500 TFs…)
- Number of conditions o = 20 (guess)
- Number of time points p = 10 (guess)
- Number of whole genome assays to fill an $m*n*o*p$ matrix = **800 million**

Focused efforts of multiple labs on one system gets closer to completeness

A.



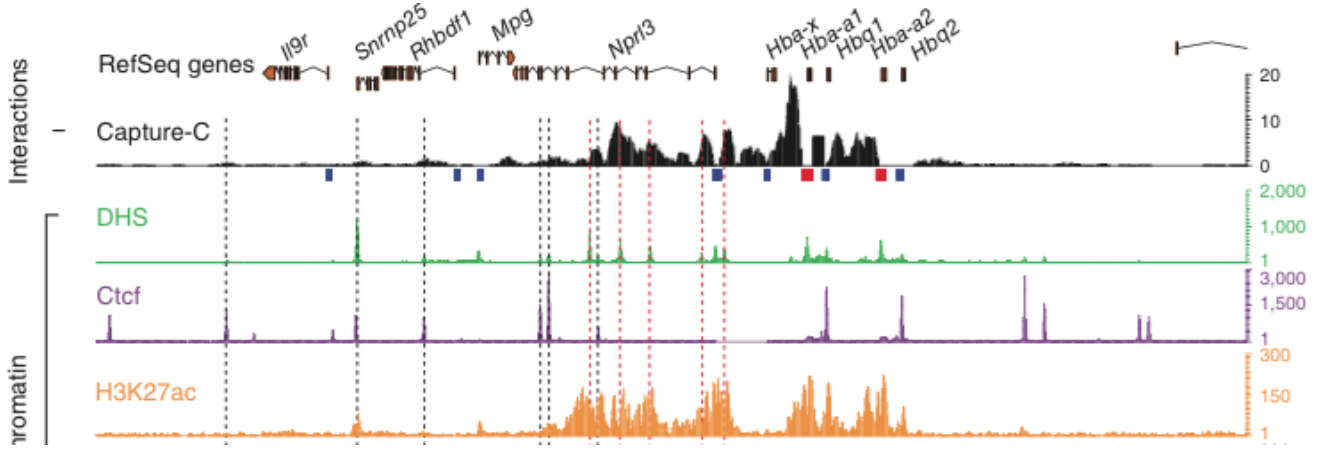
B.

	RNA		Access			DNAme		H3K						TFs	
Cell type	pA	tS	ATAC		DN	MBP	Oth	4m3	4m1	27ac	36m3	27m3	9m3	nbr	
ES_E14														4	4
HSC															5
MPP															4
HPC7															11
CMP															
MEP															
CFU-E														1	4
G1E														7	
ER4+E2														7	
ERY														3	2
CFU-M															1
MEG														4	3
GMP															
MO															
MF															
NEU															1
CLP															
B cells															
CH12														1	2

Hematopoiesis and datasets

Genome-wide datasets that are needed (1)

- 3D chromatin interaction maps



Hughes et al. (2014)
Capture C.
Nature Genetics

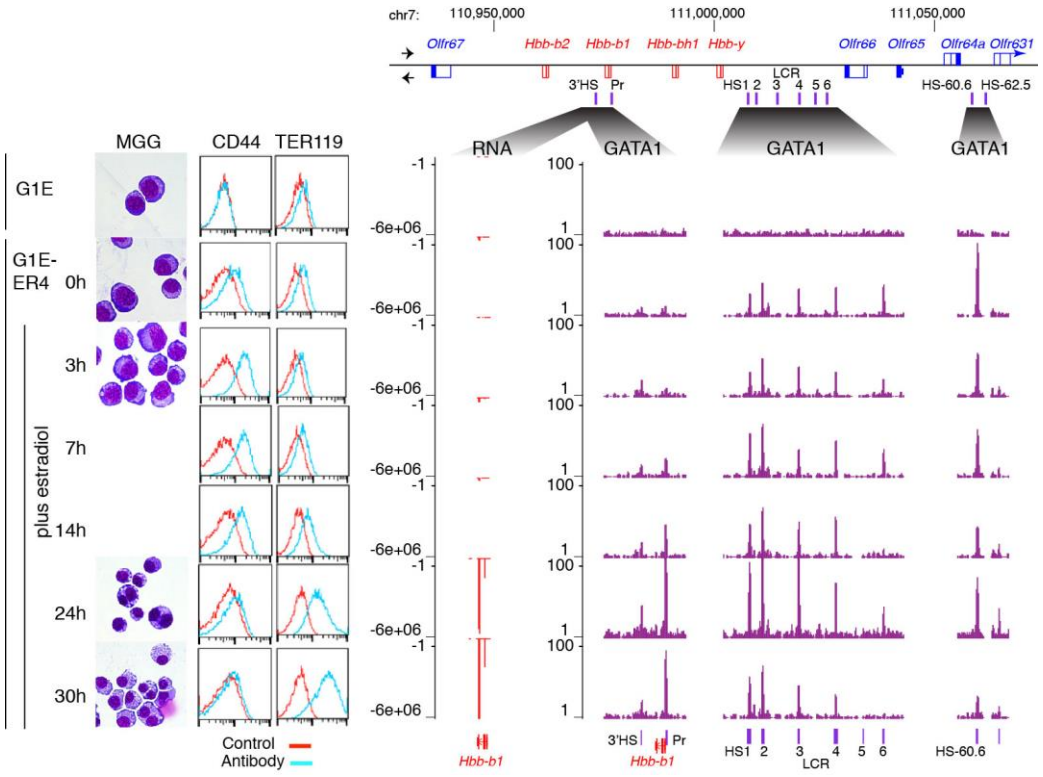
  - Range of scales (10 to 1000 kb)
  - Many cell types
- Dynamics of interaction maps
  - Across a differentiation series
  - Response to environmental stimuli
- Coordinate with and complement the 4D Nucleome project
- Top-down managed approach

Genome-wide datasets that are needed (2)

- More TFs and other features mapped in a greater number of cell types
- Current limitations:
  - ChIP/RIP-grade antibodies
  - Number of cells (10 to 20 million cells)
- Higher resolution (ChIP-exo; DNase footprints)
- Leverage DNase footprints to infer bound TF classes
- Top-down managed approach (?), Community driven (?)

Dynamics

- Follow epigenetic marks and transcriptional response across a time series in response to a stimulus – or as a normal differentiation series
- Distinguish cause from effect (causative events are earlier)
- Infer mechanism from kinetics
- Community-driven project



Jain, Mishra et al. (2015) Genomics Data

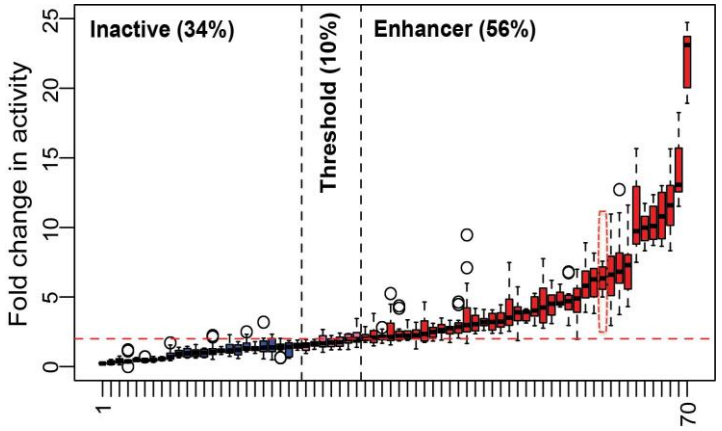
Validation and characterization of candidate functional elements (1)

- Some managed, closely coordinated efforts
- Use high throughput genetic screens/assays measuring activity of predicted functional modules and elements, e.g. regulatory modules
- Insure that a certain fraction of functional predictions from the Encyclopedia are tested
- Tested sets should include *positive* predictions.
  - Results will provide an empirical validation rate
  - Results could give insights into more precise insights into roles of the DNA segments (e.g. activity in unexpected tissue)
- Tested sets should also include *negative* predictions
  - DNA segments NOT predicted to be functional modules and elements
  - Give us an idea of the frequency we are missing things
- Only for the cell type-condition-organisms assayed.

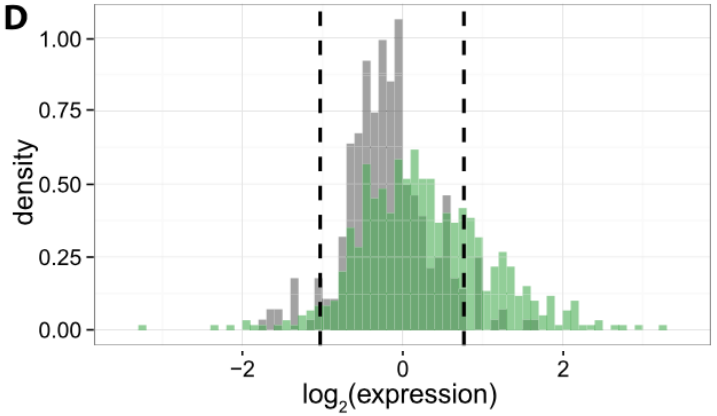
Validation and characterization of candidate functional elements (2)

- Other less tightly managed approaches
- Multiple kinds of perturbation
- Gain-of-function reporter assays
- Large-scale genetic engineering for loss-of-function and replacement mutations in the endogenous locus are critically needed.
- What aspects of this kind of work fits in the NHGRI portfolio, and what aspects belong to other Institutes?

Expand the vocabulary

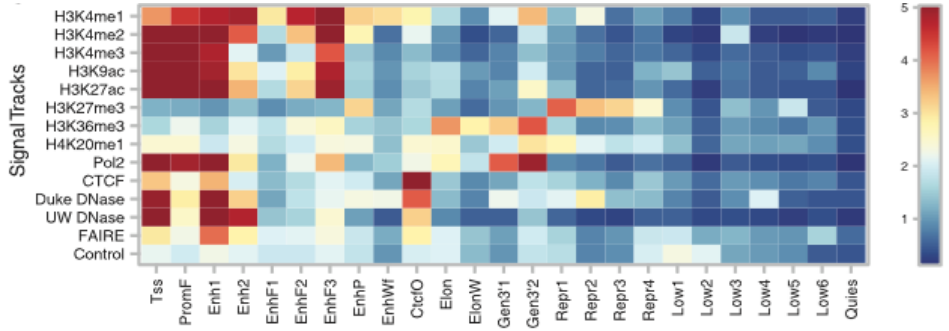


Dogan et al.submitted



Kwasnieski et al (2014) Genome Res

- Wide range of quantitative enhancer effects suggests heterogeneity, not binary classification
- Active enhancers can have diverse combinations of TFs and histone modifications
- Unsupervised learning of chromatin states suggest far more than the common 2-4 states (promoter, enhancer, silencer, insulator)



Hoffman et al (2013) Nucl AcidsRes

Encourage a wide variety of assays

- There should be periodic calls for proposals so that labs that develop a clever new assay (e.g. revealing bona fide chromatin boundaries) can be supported for genome-wide analyses

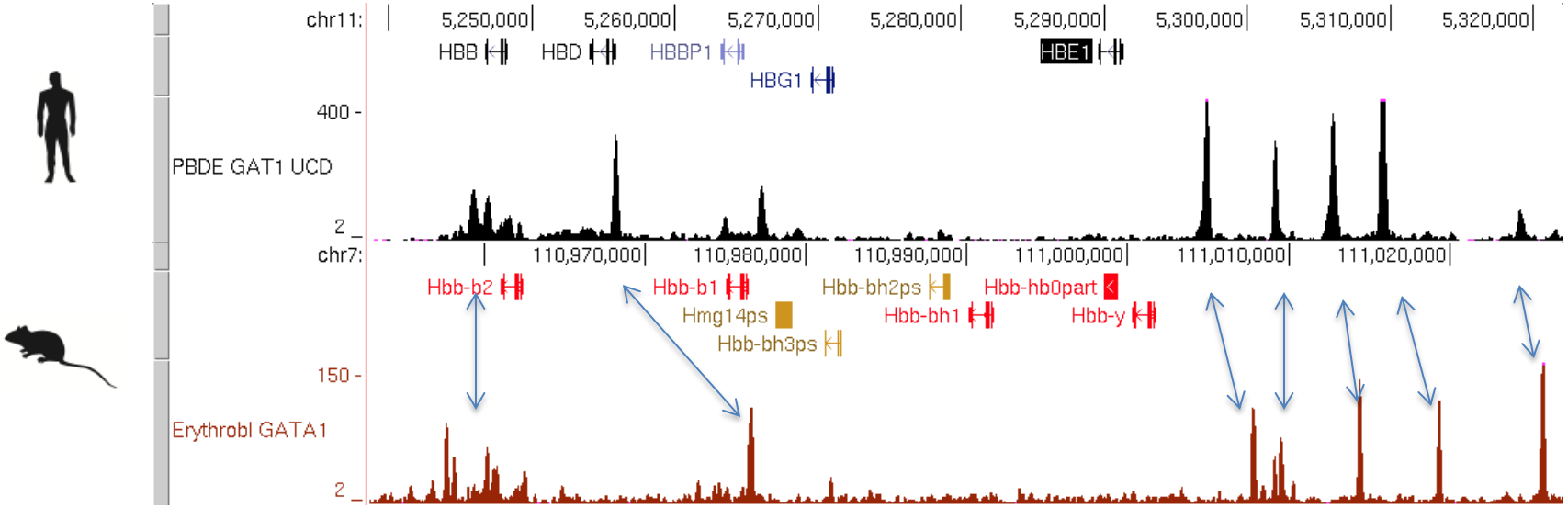
Power in interpretation of comparisons

Comparative genomics

Motifs for GATA factor binding preserved across mammals

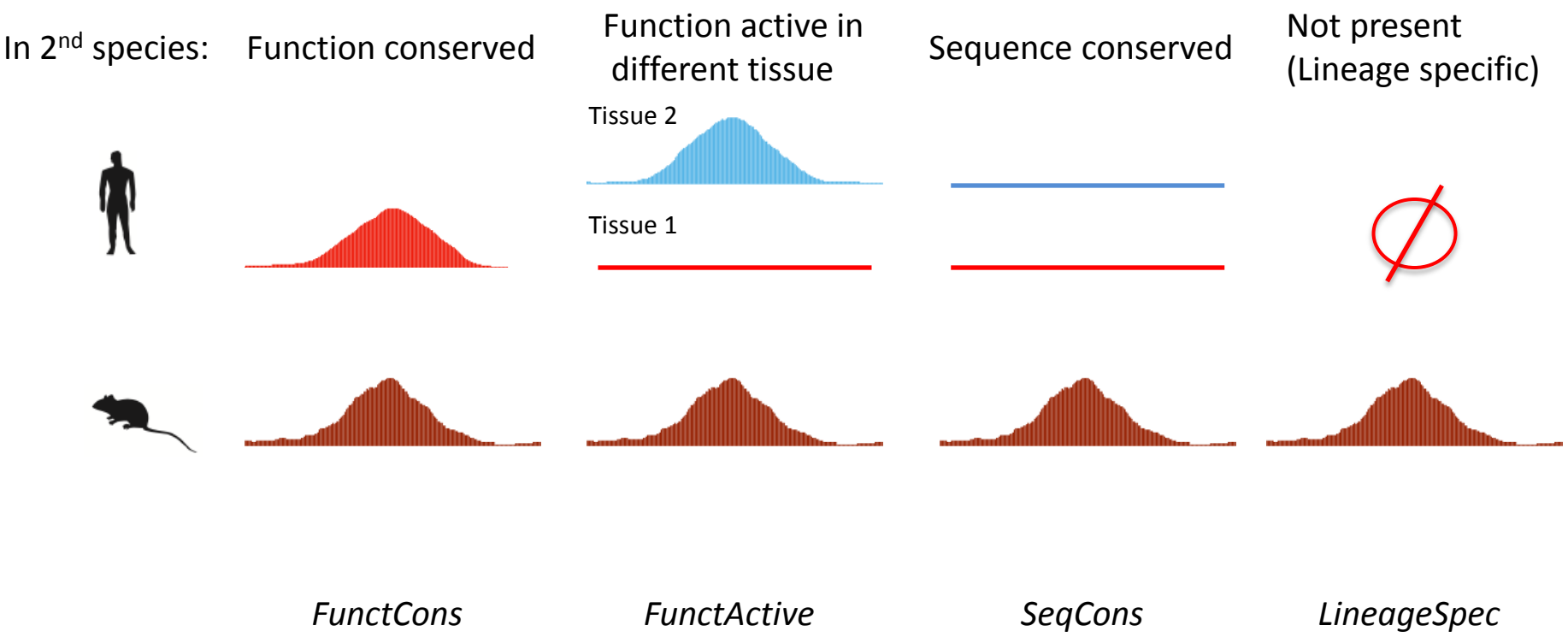
Signatures of
- purifying selection
- adaptive evolution
- lineage specificity

Comparative epigenomics



GATA1 factor occupancy in erythroblasts preserved across mammals

4 categories of functional evolution revealed by comparative epigenomics

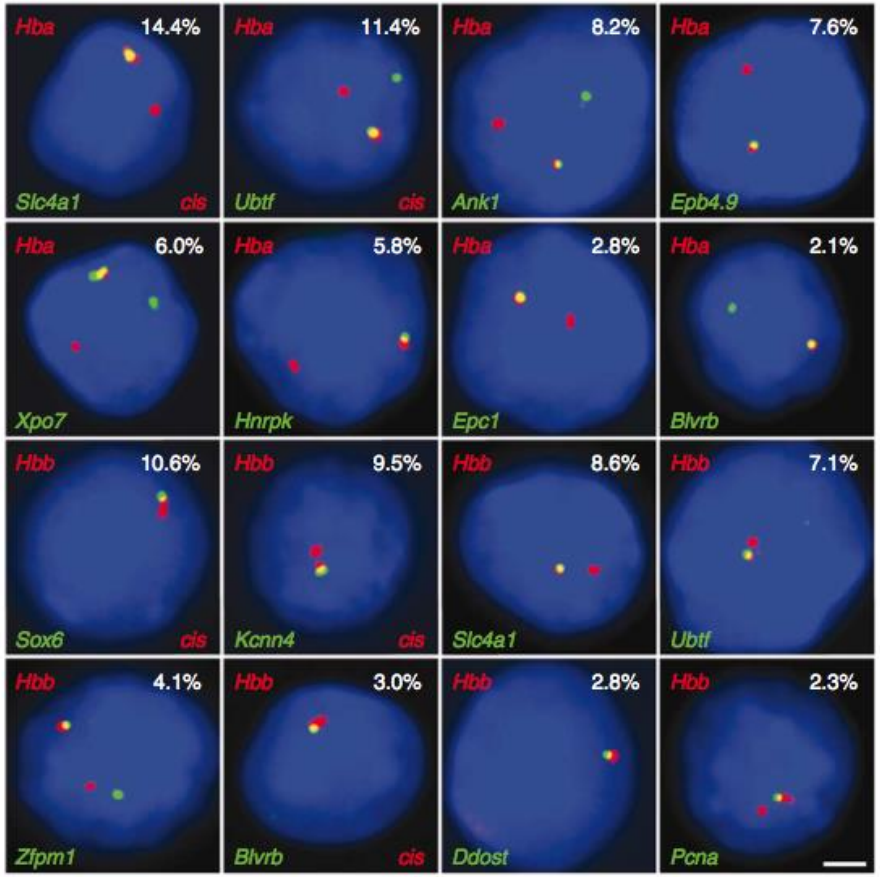


The Mouse ENCODE Consortium (2014) Nature

Denas, Sandstrom, Cheng, Beal, Herrero, Hardison, Taylor (2015) BMC Genomics; bioRxiv

What future studies could be envisaged if not limited by technology?

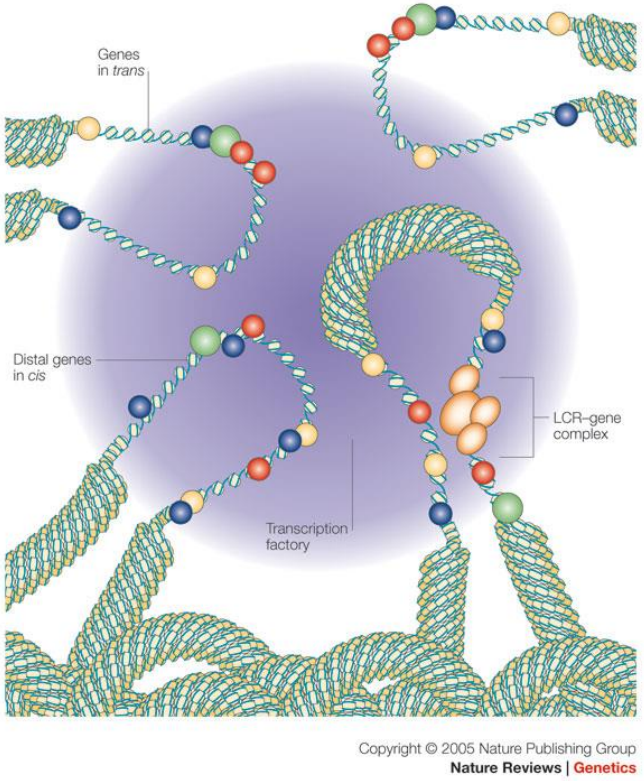
- What are the structures and mechanisms for directed movement of genes in the nucleus?
- During activation, genes move from nuclear periphery to a zone with abundant RNA POL2
  - Transcription factories?
- Actively transcribed genes co-localize



Schoenfelder et al. (2010) Nature Genetics

Directed movement

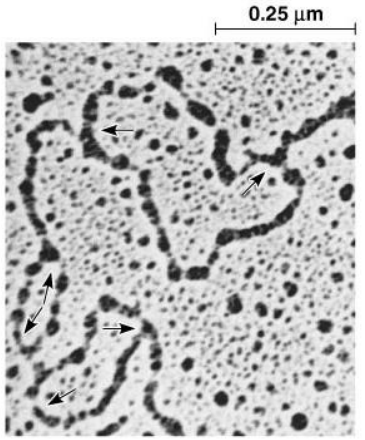
- How does a gene get to a (functionally relevant) position in the nucleus and what directed it?
- Are there molecular locomotives to pull genes along?
  - Is that what one type of enhancer does?
- Are there tracks that the gene follows?
  - Does that account for some of the unexplained TF binding?
- What determines how long a gene stays in the active zone (trancription factory?)
  - Is that what another type of enhancer does?

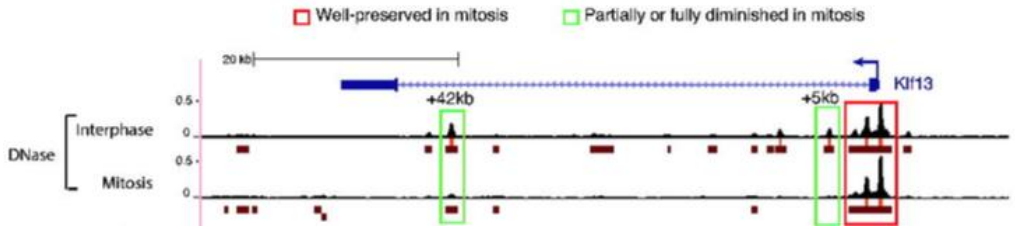


Chakalova et al. (2005) Nature Rev Genetics 6:669

Other functional elements I know I don't know

- Replication machinery and templates
  - Replication timing domains are being mapped
  - Where are the replication origins? Dynamics?
- Mitotic bookmarks
  - We can map locations that stay open and bound during mitosis



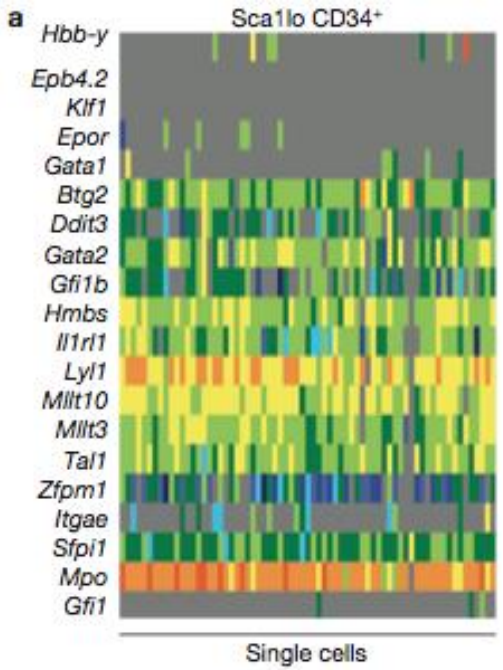


Hsiung et al (2014) Gen Res

  - What distinguishes them from other sites at which TFs dissociate during mitosis?
  - Do these sites have special roles in (re)establishing transcriptional profiles?
- Recombination hotspots
- Matrix attachment regions

Transformative technological breakthroughs (1)

- In all cases, new methods must be robust and accurate
- Mapping binding profiles for a very large number of TFs
  - Tagging TF genes by genome editing (CRISPR)
- Mapping epigenetic features on small numbers (100's to 1000's) of cells
- Transcriptomes and epigenetic profiles in single cells
  - Heterogeneity and stochastic events in single cells may reveal a radically different picture of differentiation than inferred from studies of cell populations



Pina et al. (2012) Nature Cell Biology

Transformative breakthrough (2): Visualization for interpretation

- Browsers are excellent means for studying multiple data tracks in a given locus
  - Limited to single loci
  - Limited to screen size, visual acuity – and human memory!
- Human brain can discern patterns in complex data
- Can a virtual reality viewing environment be built that would enhance integration and understanding?
- User would “fly” through a landscape representing the data (raw data, correlations, etc)

Prioritize needs

- Disease relevance is always a top priority.
- In the long run, you get a strong return on investment when the research discovers new biological insights and principles.
- Projects that dig into some newly fertile ground on enduring questions, usually in developmental biology.

What is required to make the new findings (data, computational analysis) interoperable with previous ENCODE findings? **Data coordination**

- Perhaps the most important point I can make is that any new initiative arising from these discussions has to insure
  - (a) rapid data release
  - (b) expert curation
  - (c) uniform data processing
  - (d) easy access to everyone
  - Items b, c, and d are the DCC.
- Don't think about doing anything without it.
- Continuing with previous systems is not a critical concern
- Always adopt the best methods – even if you have to drop earlier ones

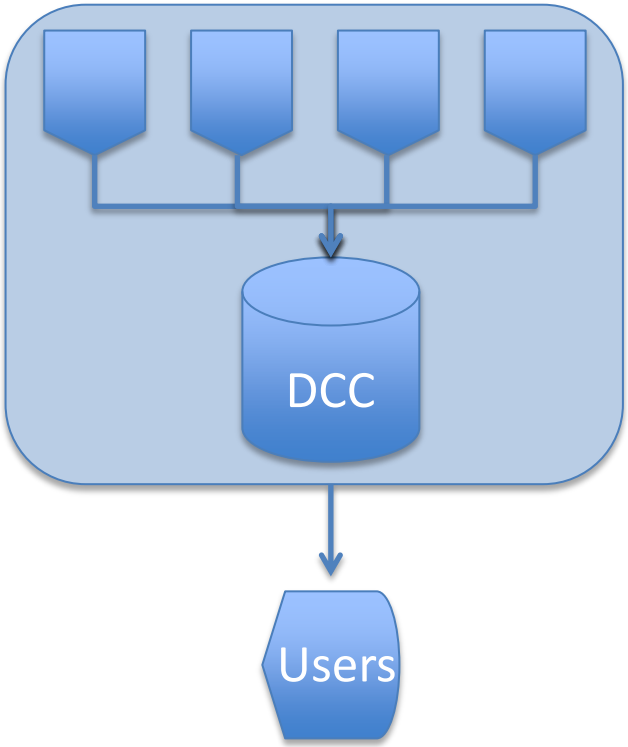
Community-driven projects

- Individual labs or groups of labs with special expertise for manipulating a system and interpreting results from genome-wide studies
- All assays are still genome-wide
- Labs contribute to and adhere to data standards
- All data are still released promptly, deposited in the Data Coordination Center

Two structures for consortium projects

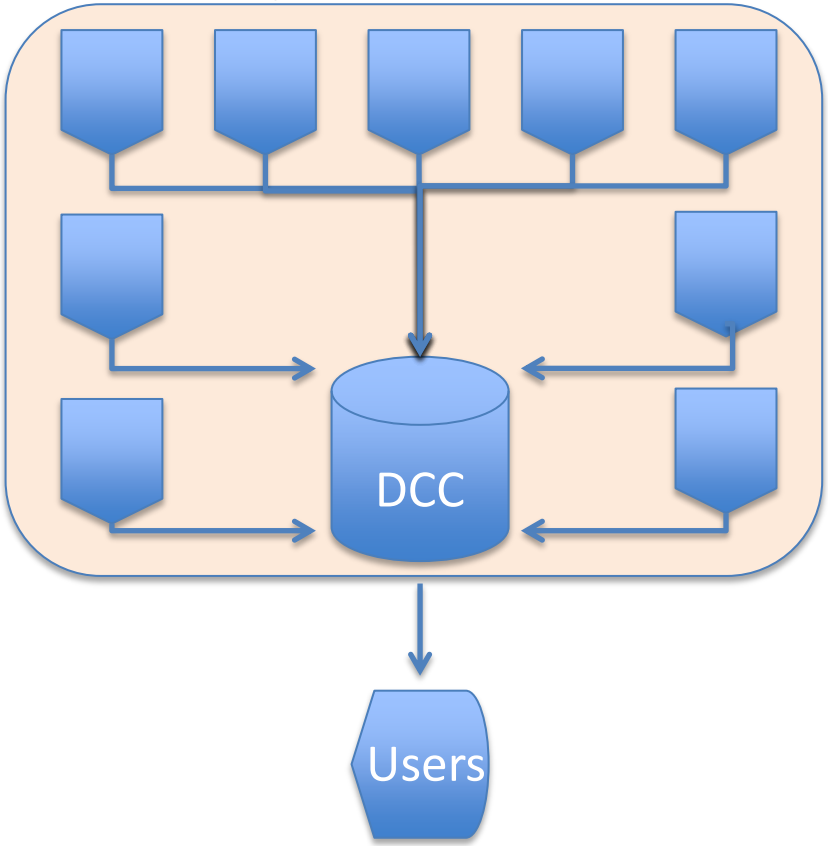
Managed, coordinated, focus on same cell types

Data production centers



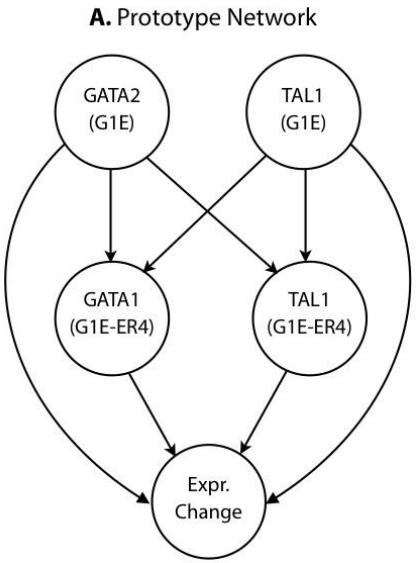
Less coordinated, systems chosen by investigators (and reviewers)

Data production centers



How do you know when information is complete?

- More TFs and other features will be assayed in more tissues, cell types, developmental stages and pathological states
- But how do we know when we know enough?
- Use predictive modeling of all epigenetic and other features to predict a (patho)physiological outcome
- See how close current knowledge leads to “understanding” = predictive accuracy
- Get more information and repeat as needed until one reaches a reasonable goal (PPV of model is high)
- Similar approach as for “Genomics of Gene Regulation”



B. Regulatory rules recovered from querying network

G1E		G1E-ER4			P(Ind) P(Rep) P(Nonresp)
GATA2	TAL1	GATA1	TAL1	Joint P	
		✓		0.4	
		✗		0.6	
✓		✓		0.06	
		✓	✓	0.2	
✗		✓	✓	0.07	
✗		✓	✓	0.03	

James Taylor JHU


Fly'in Phil's Photos
www.FlyinPhilsPhotos.com